A Bioinformatics-Based Bottom-up Network Reconstruction Approach to Detect Stem Cell-Related Blood Biomarkers of Cardiovascular Damage and Repair

Taylor Kantor[1]*, Ryan O'Neill*, Raven Edwards, Yang Wang[2], Kun Huang[2], Peter J. Mohler[1], Leni Moldovan[1], Nicanor I. Moldovan[1]**

*[1]Dorothy M. Davis Heart and Lung Research Institute, Departments of Internal Medicine and [2]Bioinformatics, College of Medicine, and Colleges of *Arts and Sciences and **Engineering Ohio State University, Columbus, Ohio, 43210*

Abstract

In peripheral blood there are very rare circulating stem/progenitor cells (CSPCs) with primitive characters and the capacity to differentiate into cardiovascular and other lineages. These cells are promising biomarkers of the organism's resilience and of its ability to repair injuries. However, because of its complexity, the system of CSPCs has not yet been described by traditional methods in a coherent and simple enough manner to be clinically useful. In response to this limitation, we proposed the CSPC system could be directly assessed by gene expression analysis of peripheral blood mononuclear cells (PBMCs). Initially, 45 genes representing the most used markers of primitivity and differentiation were tested. Among these, 15 genes were organized as a module of the blood transcriptional network which inversely depended on age, blood pressure, and vascular stiffness of donors, as expected from CSPCs. To identify more members of this module, we analyzed 503 Affymetrix microarrays from public databases hybridized with RNA of normal human PBMCs from children (where the primitive genes were better expressed), adults, and burn victims (a response to an injury condition that is preferable to actual cardiovascular patients because of the lack of other risk factors which complicate the interpretation). Normalized data was analyzed by the bioinformatics co-variation method known as "guilt-by-association". This approach identified 107 potential candidates from data collected from microarrays on healthy children, many having known roles associated with stemness, differentiation, angiogenesis, and/or cardiovascular diseases or repair. A larger study on adults was conducted, as well as studies involving burn injury (that induces massive mobilization of CSPCs) and pregnant women suffering from preeclampsia (a condition due to deficiencies in CSPCs), for comparison. In conclusion, we show how to expand a new, collective, systemic biomarker for the elusive CSPCs that could be used to track their response to injury and to identify patients at risk for developing, or already experiencing, cardiovascular diseases.

Results

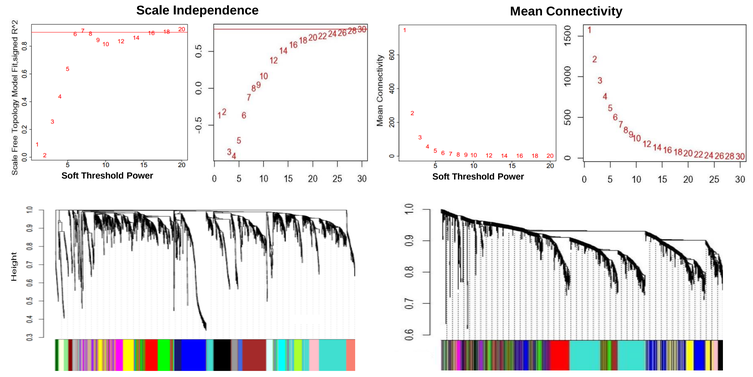


Figure 2. WGCNA attempt to identify additional CVR candidate genes. Scale-Free Topology Graph, Mean Connectivity, and Cluster analysis of genes co-varying with at least three CVR members performed by Weighted Gene Covariation Network Analysis (WGCNA, Langfelder P, Horvath S.BMC Bioinform. 2008;9:559). This was done using 8,000 of the most variable genes, a listing of over 300 stem and progenitor cell related genes suggested by the NIH, as well as 2,755 genes gathered from the GBA method which were correlated (R > 0.4) with at least 3 of the 5 present CVR genes (Presence > 0.5). Left side represents typical results of WGCNA method while the right side represents results from our GBA analysis using WGCNA. Note that higher beta values lead to a loss of information as the Pearson correlations approach zero. Higher mean connectivity values as well as higher beta values required to reach Scale Free Topology would not allow for this analysis to be used for stemness-related genes in peripheral blood.

Genes with Large Fold Change				Burn Study		Preeclampsia Study			
Entrez	Affymetrix ID	Gene Symbol	Function	1-10 Day (Early) Fold Change (Compared to Controls)	11-49 Day (Late) Fold Change (Compared to Controls)	11-49 Day Fold Change (Compared to Early)	Early onset Fold Change (Compared to Controls)	p-value	Late onset Fold Change (Compared to Controls)	p-value
4317	207329_at	MMP8	Plays an important role in plaque angiogenesis	**2.63**	**3.32**	0.68	-	-	1.99	0.037
4057	202018_s_at	LTF	Highly expressed in non-healing wound exudates	**2.36**	**3.09**	0.73	-	-	2.03	0.001
3934	212531_at	LCN2	An independent renal predictor of incidence of AKI after surgical abdominal aortic aneurysm (AAA) repair	**2.30**	**3.20**	0.89	1.58	0.098	1.85	0.001
10562	212768_s_at	OLFM4	Marker of intestinal stem cells	**1.96**	**2.39**	0.43	-	-	1.88	0.077
10321	207802_at	CRISP3	Specific neutrophil granules protein, secreted as an extracellular matrix component	**1.51**	**1.12**	-0.39	-	-	1.83	0.052
671	205557_at	BPI	Antibacterial activity against Gram-negative bacteria	**1.43**	**2.43**	**1.00**	2	0.021	2.01	0.005
1669	207269_at	DEFA4	Neutrophil protein with antimicrobial activity against Gram-negative bacteria	0.55	**2.66**	**2.11**	2.36	0.011	2.49	0.001
4680	211657_at	CEACAM6	Implicated in cell adhesion, cellular invasiveness, angiogenesis, and inflammation	0.45	**2.42**	**1.96**	2.1	0.041	1.99	0.023
6037	206851_at	RNASE3	Eosinophil major basic protein with pro-angiogenic effects	0.54	**1.42**	0.89	-	-	1.89	0.027
154664	1553605_s_at	ABCA13	Markers of hMSC	0.18	**1.16**	0.98	1.56	0.086	-	-
820	210244_at	CAMP	Induces angiogenesis via PGE2-EP3 signaling in endothelial cells	0.59	**1.06**	0.47	-	-	1.66	0.002
4318	203936_s_at	MMP9	Involved in mobilization of hematopoietic progenitor cells from bone marrow as well as embryonic development, reproduction, and tissue remodeling through ECM degradation	**2.56**	**2.56**	-0.01	1.62	0.039	-	-
8993	207384_at	PGLYRP1	Binds to peptidoglycan of bacteria involved in human atherosclerotic lesions. Increasing levels associated with coronary artery calcification and aortic wall thickness	**1.69**	**2.12**	0.43	1.59	0.007	-	-
762	206209_s_at	CA4	Encodes an isozyme expressed on luminal surfaces of pulmonary and other capillaries and proximal renal tubules. Related to Ocular Hypotension	**1.54**	**1.37**	-0.17	1.52	0.040	-	-
660	206464_at	BMX	Plays a critical role in TNF-induced angiogenesis, and implicated in the signaling of TEK and FLT1 receptors, 2 important receptor families essential for angiogenesis	**1.54**	**1.61**	0.07	1.5	0.033	-	-
56729	220570_at	RETN	Resistin levels correlate with oxidative stress and myocardial injury in cardiac surgery patients. It may serve as a useful biomarker for ischemia-reperfusion injury	**1.53**	**1.63**	0.09	-	-	1.53	0.060
6947	205513_at	TCN1	Influences human vitamin B-12 levels which have shown to be implicated in congenital heart disease via the folate metabolism pathway	**1.47**	**2.29**	0.82	1.75	0.012	1.68	0.008
1088	206676_at	CEACAM8	Leukocyte activation marker. Increased leukocyte activation is correlated with coronary artery disease, plaque destabilization, and vascular cell dysfunction	**1.18**	**2.86**	**1.68**	1.9	0.051	2.09	0.004
1232	208304_at	CCR3	CCR3-dependent chemokine interactions regulate endogenous migration of CD34+ progenitors from bone marrow to ischemic but not to normal myocardium	-1.01	-0.90	0.10	1.84	0.000	1.64	0.001
260429	1552348_at	PRSS33	Serine, protease predominantly expressed in macrophages	-1.02	-0.47	0.55	-	-	1.73	0.087
100133941	216379_x_at	CD24	Interacts with P-selectin which mediates rapid rolling of leukocytes over vascular surfaces during the initial steps in inflammation	0.70	**1.90**	**1.19**	1.63	0.039	1.52	0.033
932	210254_at	MS4A3	Hematopoietic stem cell cycle regulator	-0.05	**1.79**	**1.84**	1.83	0.096	1.63	0.085
5806	206157_at	PTX3	Modulates inflammatory processes, angiogenesis, atherosclerotic lesion development, and ECM formation. Released by vascular wall cells as an inflammatory marker	0.44	**1.18**	0.74	1.57	0.045	1.69	0.002
983	206177_s_at	ARG1	Reduces nitric oxide production and impairs endothelial function	**2.68**	**2.74**	0.05	-	-	-	-
306	209369_at	ANXA3	Potential angiogenic mediator	**2.21**	**2.31**	0.10	-	-	-	-
5004/5005	205041_s_at	ORMU2	Acute phase reactant and bimodal regulator of angiogenesis	**1.00**	**1.06**	0.06	-	-	-	-
4353	203949_at	MPO	Leukocyte activation marker. Increased leukocyte activation is correlated with coronary artery disease, plaque destabilization, and vascular cell dysfunction	0.39	**1.61**	**1.22**	-	-	-	-
5004	205040_at	ORM1	Acute phase reactant and bimodal regulator of angiogenesis	0.98	**1.16**	0.18	-	-	-	-
4116	209395_at	CHI3L1	Expressed by macrophages, chondrocytes, and vascular SMC; it is a potent angiogenic factor	-0.97	0.03	**1.00**	-	-	-	-

Table 2. Genes with the largest fold change in burn and preeclampsia patients, as compared to prevalence in children vs. adults. Orange = Prevalent in burn victims, preeclampsia, and present in the children's network; Yellow = Prevalent in burn victims and preeclampsia, but not present in children; Red = Prevalent in burn victims and present in the children's network, but not prevalent in preeclampsia.

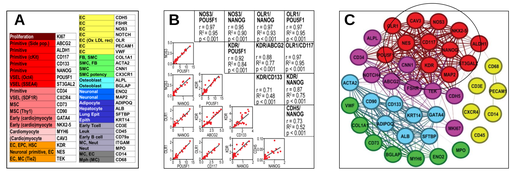


Figure 1. Defining a core network module with cardiovascular function ('cardiovascular repairome', CVR) by PCR analysis (Moldovan et al., *PlosONE*, in press). **A**. Panel of primitive (left column) and differentiation (right column) tested genes. **B**. Example of covariation of primitive and cardiovascular genes in PBMCs from 27 healthy human subjects. **C.** Covariation expression matrix of genes in PBMCs from normal subjects. A module composed of 15 genes called CVR (encircled) were found to inversely relate to blood donor's age, vascular stiffness and central blood pressure parameters, suggestive of a vascular protective role similar to CSPCs. The goal of the current project is to expand this network using the Guilt-by-Association (GBA) principle.

Table 1: Poor sensitivity of microarrays for detecting genes with low transcripts. Affymetrix microarrays from listed GSE studies were analyzed. The table contains 'presence scores' on Affymetrix GeneChips® of the selected gene panel (see Fig. 1). Highlighted (yellow) are members of the CVR module. Stars (**) represent genes whose presence levels were higher in children than adults. Most of the CVR genes were not present in a study where CD34+ cells were pre-selected (of note, even CD34 had an incomplete representation). Also indicated are MKI67 and OLR1, two genes with increased presence in samples from burn injury.

Table 1. Detection of select primitive and differentiation genes on Affymetrix GeneChip microarrays.

Affymetrix ID[*1]	Gene Symbol	RefSeq Transcript ID[*2]	Presence score[*3] Children[*4]	Adult[*5]	Early Burn	Late Burn	CD34[*6]+
208264_s_at	CAV2	NM_001233	0%	0%	0%	0%	0%
1570276_a_at	GATA4	NM_002052	0%	0%	0%	0%	0%
209351_at	KRT14	NM_000526	0%	0%	0%	0%	0%
218678_at	NES	NM_006617	0%	0%	0%	0%	0%
206578_at	NKX2-5	NM_001166175	0%	0%	0%	0%	0%
217711_at	TEK	NM_000459	0%	0%	0%	0%	0%
213869_x_at	THY1	NM_006288	0%	0%	2%	0%	0%
207175_at	ADIPOQ	NM_001177800	0%	15%	11%	16%	25%
209543_s_at	CD34	NM_001025109	0%	9%	0%	2%	88%
211201_at	FSHR	NM_000145	2%	0%	0%	0%	0%
214468_at	MYH9	NM_002473	2%	5%	0%	0%	0%
203951_at	CNN1	NM_001299	2%	1%	0%	2%	9%
202112_at	VWF	NM_000552	2%	3%	0%	7%	71%
214837_at	ALB	NM_000477	3%	3%	4%	4%	4%
217430_x_at	COL1A1	NM_000088	7%	32%	4%	2%	39%
204677_at	CDH5	NM_001114117	10%	0%	0%	0%	2%
214532_x_at	POU5F1**	NM_001159542	12%	1%	0%	2%	0%
203934_at	KDR	NM_002253	12%	16%	12%	5%	16%
215783_s_at	ALPL	NM_000478	15%	82%	100%	98%	36%
209735_at	ABCG2**	NM_001257386	24%	10%	9%	11%	73%
205051_s_at	KIT**	NM_000222	44%	24%	12%	7%	98%
204304_s_at	PROM1**	NM_001145847	61%	28%	21%	42%	100%
208247_at	NOTCH4	NM_004557	61%	67%	54%	42%	64%
212022_s_at	MKI67**	NM_001145966	64%	1%	23%	65%	18%
210004_at	OLR1	NM_001172632	64%	64%	82%	100%	89%
225540_at	MAP2**	NM_001039538	78%	8%	16%	19%	21%
220184_at	NANOG**	NM_024865	83%	0%	0%	2%	11%
203949_at	MPO**	NM_000250	83%	24%	75%	93%	95%
200974_at	ACTA2	NM_001141945	92%	99%	98%	100%	98%
229093_s_at	NOS3**	NM_000603	93%	0%	0%	0%	0%
206956_at	BGLAP	NM_001199661	93%	92%	79%	81%	84%
214354_x_at	SFTPB**	NM_000542	98%	21%	7%	11%	13%
203507_at	CD68**	NM_001040059	98%	84%	82%	74%	39%
201313_at	ENO2**	NM_001975	100%	70%	25%	19%	91%
217650_x_at	ST3GAL2**	NM_006927	100%	84%	88%	91%	96%
212224_at	ALDH1A1	NM_000689	100%	93%	39%	33%	96%
205456_at	CD3E	NM_000733	100%	99%	93%	93%	25%
1555779_a_at	CD79A	NM_001783	100%	99%	91%	86%	38%
203939_at	NT5E	NM_001204813	100%	99%	81%	65%	55%
205898_at	CX3CR1	NM_001171171	100%	100%	100%	100%	61%
201743_at	CD14	NM_000591	100%	100%	96%	96%	88%
205786_s_at	ITGAM	NM_000632	100%	100%	100%	100%	96%
217028_at	CXCR4	NM_001008540	100%	100%	100%	100%	100%
208982_at	PECAM1	NM_000442	100%	100%	100%	100%	100%
212587_s_at	PTPRC	NM_001267798	100%	100%	100%	100%	100%
201891_s_at	B2M	NM_004048	100%	100%	100%	100%	100%
212191_x_at	RPL13	NM_000977	100%	100%	100%	96%	100%
213798_s_at	CAP1	NM_001105530	100%	100%	100%	100%	100%
217398_x_at	GAPDH	NM_001256799	100%	100%	100%	100%	100%

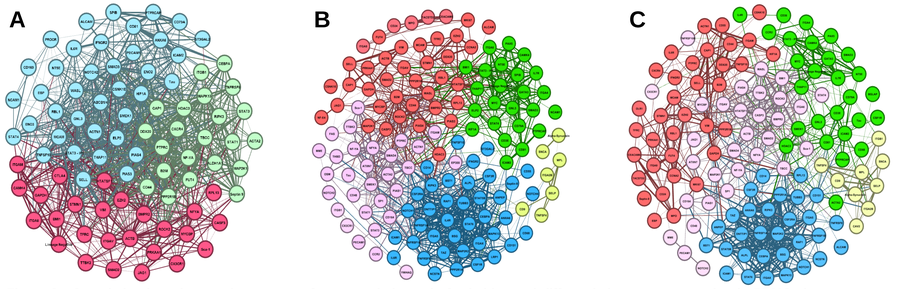


Figure 3: Covariation-based network structure of an extended panel of primitive and differentiation genes occurring in PBMCs in response to burn injury (extended gene panel compiled by NIH's *Stem Cells Interest Group*)**.** Network detection of genes with a Presence Score > 0.5 via Clique Mining was performed on over 300 suggested stem and progenitor cell related genes. Networks were created for controls **(A)**, early burn victims (1-10 days) **(B)**, and late burn victims (11-49 days) **(C)**. **A**. No modules were common between the controls and burn victims as is established by the different colored modules. **B-C**. However, similar modules did arise between the early and late burn victims. Two modules had a majority of their members present in both groups (blue and yellow) while the red and green modules had more limited similarities between groups.

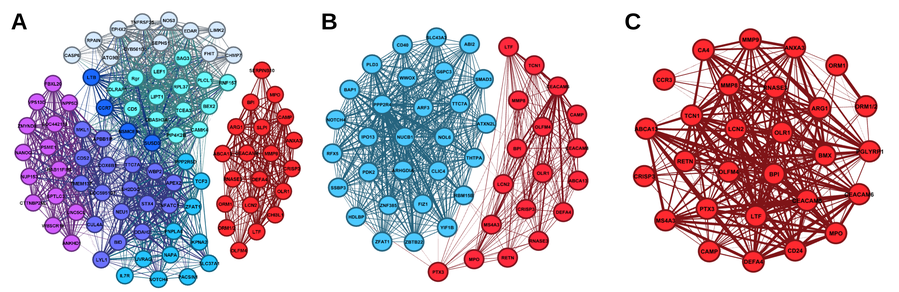
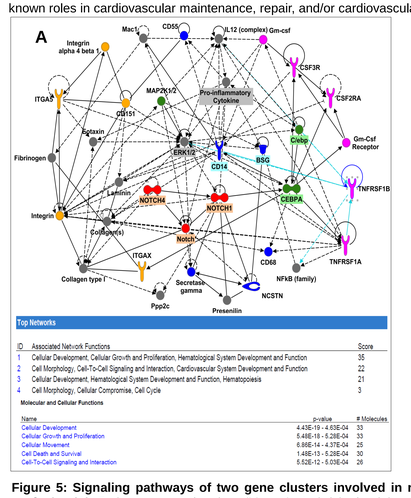


Figure 4: Network reconstruction from the covariation 'neighborhoods' of CVR seed genes. Network creation and module detection was done by performing regression analysis followed by a Clique Mining visualization on Affymetrix GeneChips® based on RNA isolated from PBMCs of healthy children **(A)**, healthy adults **(B)**, and burn victims **(C)**. 'Seed' genes were required to have a presence score of at least 50%. **(A)** The red module consists of the seed gene OLR1 and its most highly correlated 'neighbors'. Additional seed genes, NOTCH4 (bottom), NANOG (left), and NOS3 (top), are also represented along with their most highly correlated neighbors (Pearson correlation > 0.7). Genes on the gradient of purple to blue are highly interconnected with one another as established by the large density of edges between them. OLR1's neighborhood has no connections with the others at this edge threshold (w = 0.6). **(B)** Colors corresponding to NOTCH4's and OLR1's neighborhoods were kept consistent as NOTCH4 and OLR1 were the only seed genes used in the adult analysis (Presence Score > 0.5). While most of NOTCH4's members are vastly dissimilar from the Children's Network, many of OLR1's neighbors remained some of the most highly correlated genes in the adult study. Edge strength was lowered (w = 0.4) to show the consistent limited connections between the two modules. **(C)**. Genes exhibiting the largest fold change in expression value from controls to early and late term burn victims were recorded. These were compared to genes which had high fold change in expression level from the preeclampsia study. Genes with large fold changes (>1 on log scale) in both studies, as well as genes with large fold changes that were represented in OLR1's network in children, were kept for analysis. Module representation shows much higher overall connectivity and strength in connectivity in OLR1's module, though OLR1 is slightly less connected.

Seed Gene	ID	Gene Symbol	Gene Name	R	Functional Relevance
NOS3	221790_s_at	LDLRAP1	Low density lipoprotein receptor adaptor protein 1	0.798	Linked to hypercholesterolemia through mutations which prevent interaction with the LDL receptor
	209368_at	EPHX2	Epoxide hydrolase 2, cytoplasmic	0.780	Regulates vascular tone, nociception, angiogenesis and inflammation. Central role in cardiac hypertrophy, diabetes, hypertension, and pain
	210349_at	CAMK4	Calcium/calmodulin-dependent protein kinase IV	0.765	Indicated in hypertension and regulation of vascular tone. Plays a pivotal role in BP regulation through the control of endothelial nitric oxide synthase activity
NANOG	232466_at	CUL4A	Cullin 4A	0.757	Required for hematopoietic stem-cell engraftment and self-renewal
	1568943_at	INPP5D	Inositol polyphosphate-5-phosphatase, 145kDa	0.753	Expression is restricted to hematopoietic cells. The protein functions as a negative regulator of myeloid cell proliferation and survival
	1559902_at	MKL1	Megakaryoblastic leukemia (translocation) 1	0.704	Regulates myofibroblast activation and fibrosis in response to the renin-angiotensin system and post-MI remodeling
NOTCH4	214909_s_at	DDAH2	Dimethylarginine / Dimethylaminohydrolase 2	0.737	Expression promotes angiogenesis via regulating VEGF-dependent pathway. Also plays a role in NO generation through methylarginine regulation
	226564_at	ZFAT1	ZFAT zinc finger 1	0.727	Essential for endothelial cell assembly, and may play a critical role in the capillary-like network formation that is involved in vascular remodeling
	203243_at	UVRAG	UV radiation resistance associated gene	0.725	Epithelial organizer which directly impinges on epidermal-mesenchymal transition, stemness, senescence, cell death, and cell cycle arrest

Table 3: Most highly correlated 'neighbors' to CVR seed genes found by GBA in children: Examples of candidate genes for repairome extension are listed here. These genes have been found to be highly correlated with one of the original CVR genes and all have known roles in cardiovascular maintenance, repair, and/or cardiovascular diseases.



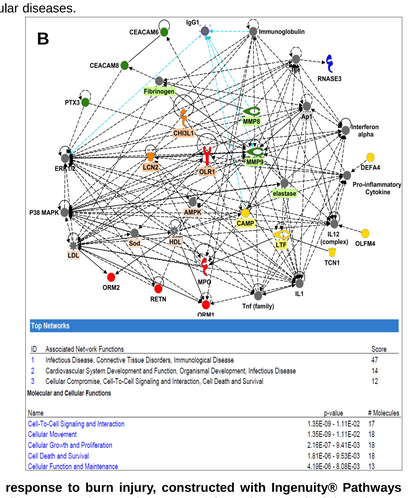


Figure 5: Signaling pathways of two gene clusters involved in response to burn injury, constructed with Ingenuity® Pathways analysis. (A) Pathways associated to 'blue cluster' (obtained by direct targeted network representation, Fig. 3 B,C). **(B)** Pathways associated to 'red cluster' obtained by GBA method. Of note, although the composition is different, both address similar functions and both are controlled by the transcription factor CCAAT/enhancer-binding protein alpha (CEBPA) (Rosmarin et al., Exp Hematol. 2005 Feb;33(2):131-43)

Materials and Methods

- *Subjects*: 59 children (Barnes et al., Arthritis Rheum. 2009 Jul;60(7):2102-12), 274 adults from 7 studies, 56 people from CD34+ study (Li et al., Cancer Cell. 2011 Nov 15;20(5):591-605), and 57 burn victims over two time points (early and late) (Xu et al., Proc Natl Acad Sci U S A. 2010 Jun 1;107(22):9923-8).
- Microarray data files (GSMs) were downloaded from PubMed studies: Children: GSE13501(59); Adult: GSE21942 (15), GSE27034 (18), GSE14642 (20), GSE11761 (20), GSE46480 (98), GSE8507 (17), GSE10041 (23); CD34+: GSE23025 (56). Only controls were used. GSE19743 (63-controls, 57 early/late burn victims).
- *Microarrays:*
  - Microarrays were Affymetrix GeneChip® Human Genome U133 Plus 2.0 Arrays (GPL570)
- *Data Processing Methods:*
  - Single-Channel Array Normalization (SCAN)
  - Presence / Absence Scores using Affymetrix Expression Console
  - Pearson Correlation and Matrix analysis using Partek Discovery Suite
  - Weighted Gene Correlation Network Analysis (WGCNA) using the R program.
  - Clique Mining Network Detection

Conclusions

1. **In this proof of concept study, we demonstrate how a 'guilt-by-association' principle-based bottom-up approach can be applied to reconstruct the transcriptional network, and thus to obtain new candidate genes as members of a pre-existing gene module containing rare primitive and differentiation markers, in spite of the low sensitivity of microarrays for mRNA detection.**
2. **Using this new method, we found a cluster of genes rich in primitive markers, expressed in PBMCs as co-variants of OLR1 (oxidized LDL receptor, LOX-1), that recurrently occurs on Affymetrix microarrays from normal children and adults, or in response to burn injury, and also being among the most modified genes in women with preeclampsia.**
3. **At the same time, the coordinate reduction - or going undetectable - of other gene clusters during physiological or pathological perturbations illustrate the complex response to injury and repair of the transcriptional landscape in peripheral blood, with contribution from the release into the circulation and recruitment of progenitor cells, and from modulation of gene expression within these cells.**
4. **The large participation of detected modules of common signaling pathways, as well as their common putative control by transcription factors, retroactively brings credit to the hypothesis that expression co-variation is derived from functional linking (i.e. 'the guilt-by-association' principle).**

Acknowledgements

This work was supported by the NIH grants RC2 AG-036559 and R01 HL-65983 (to NIM) and by a Research Fellowship award from The Ohio State University's Arts and Sciences Honors Committee to Taylor Kantor.